A Framework for Scalable Trainable Image-based Query in Video

Jianbo Shi

Informedia-II
Carnegie Mellon University

Informedia Project

- Multi-modal query of video
  - Speech recognition
  - Text analysis
  - Object detection in video
- Visualization and Summarization
  - Multiple Video Documents
  - Topic collages

http://www.informedia.cs.cmu.edu

Lesson learned

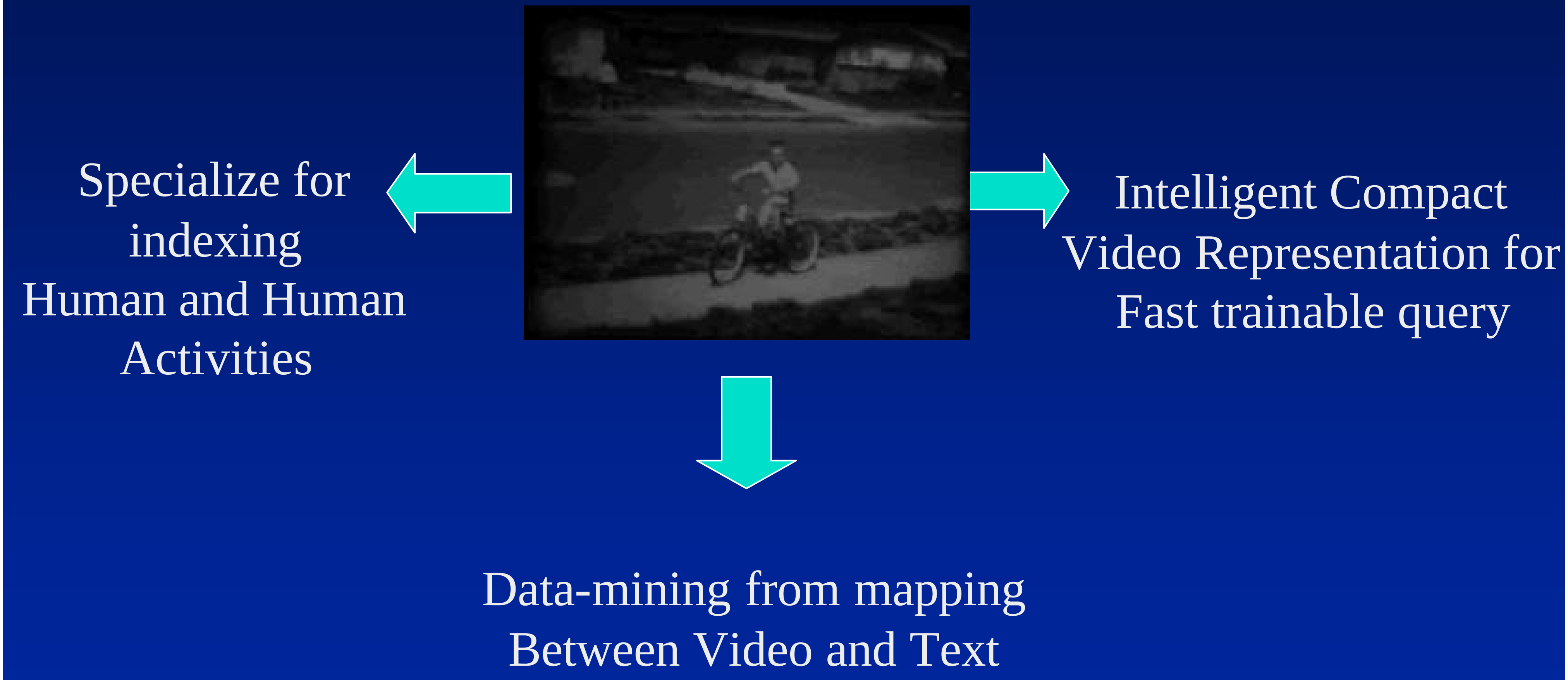
- Image-based query is a powerful tool for finding relevant information
  - Informedia-I, Blobworld, QBIC, ect.

Lesson learned

- Image-based query is a powerful tool for finding relevant information
  - Informedia-I, Blobworld, QBIC, ect.

- …but we are far from achieving
  - Human level object recognition
  - Rapid processing of massive video/image data
    - *The story is in the large of volume data*

Scaling Video Image Processing Engine



Specialize for indexing Human and Human Activities

Intelligent Compact Video Representation for Fast trainable query

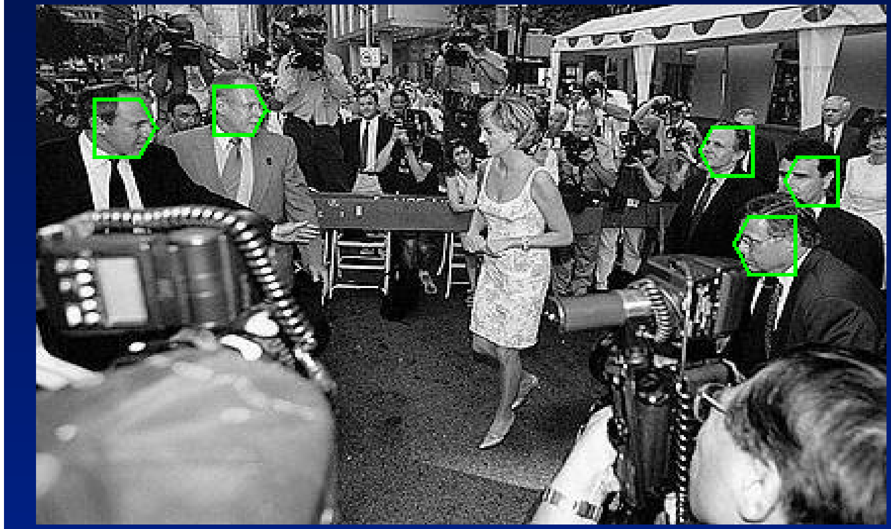
Data-mining from mapping Between Video and Text

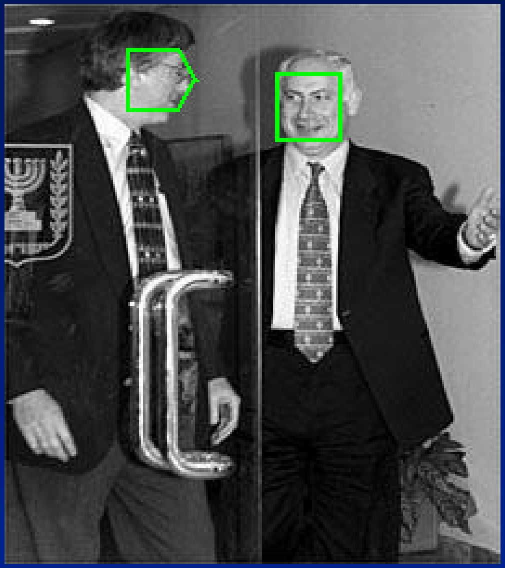
informedia

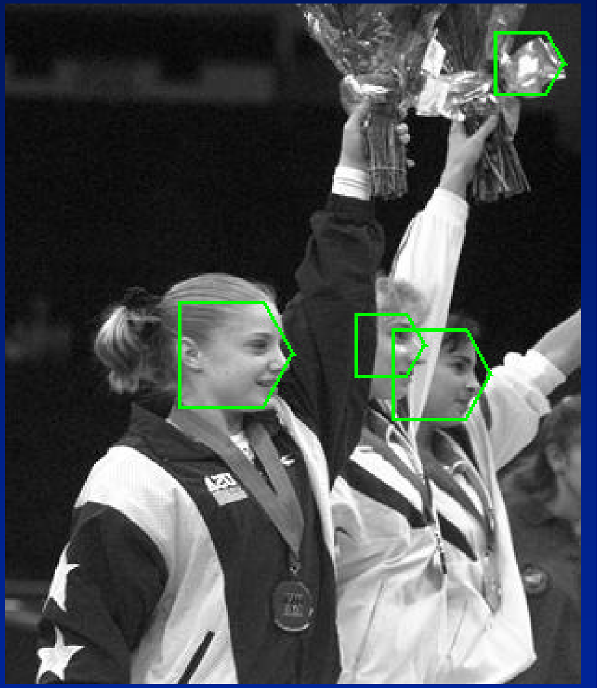
Human interests

Human Identity: Face

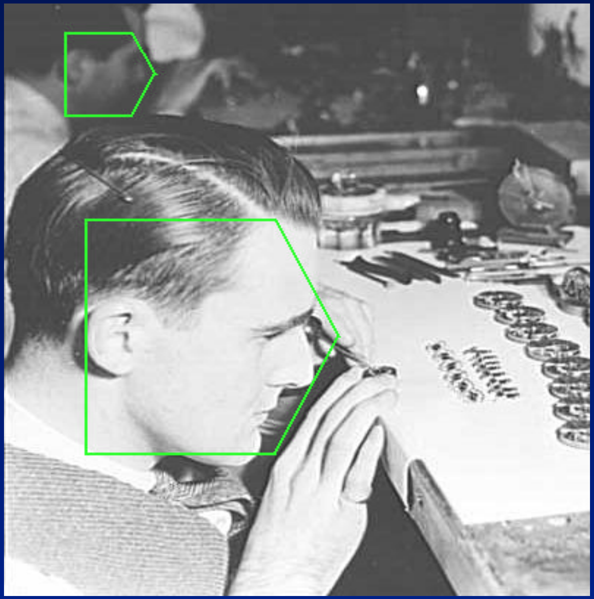
Human Activity Recognition: Body

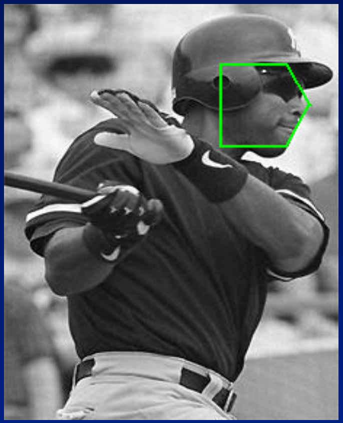


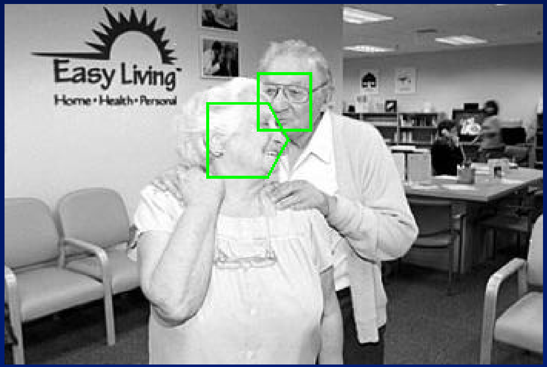


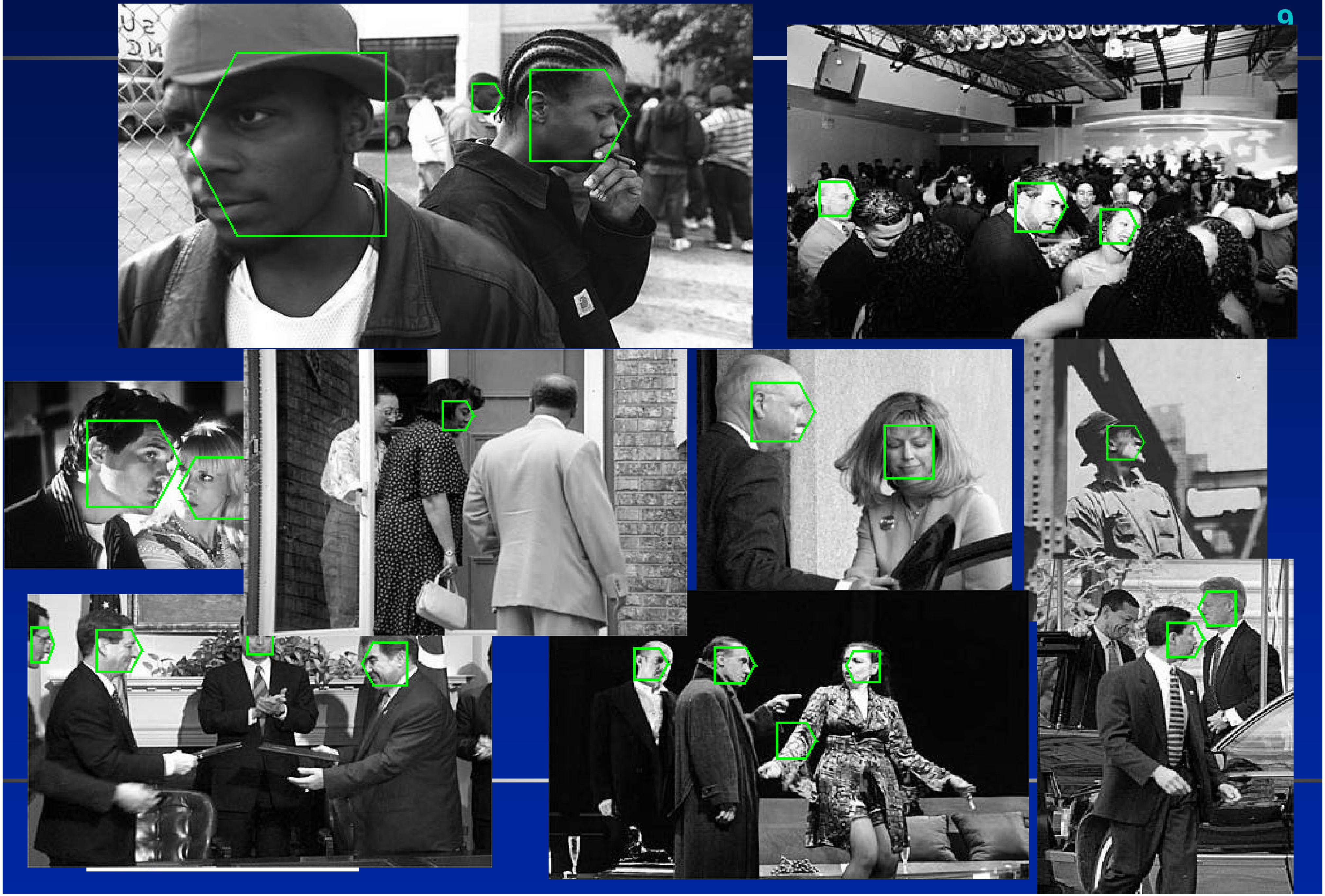


Face Detection - Henry Schneiderman (CMU)






Easy Living
Home • Health • Personal



Body Detection/Activity Recognition in Video

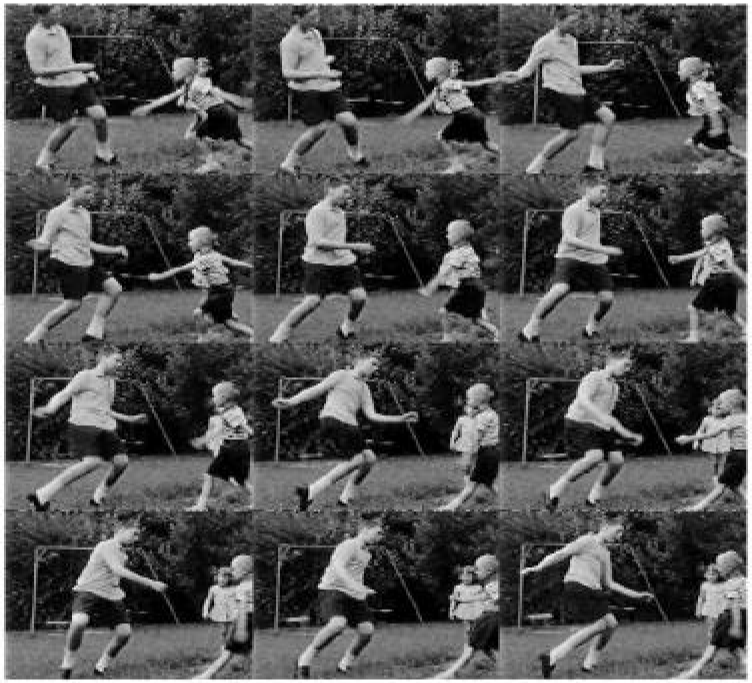
Detecting detecting human in video

Understanding human activity

Measuring/analyzing body movement

Recognizing human from body movement

The main difficulty:



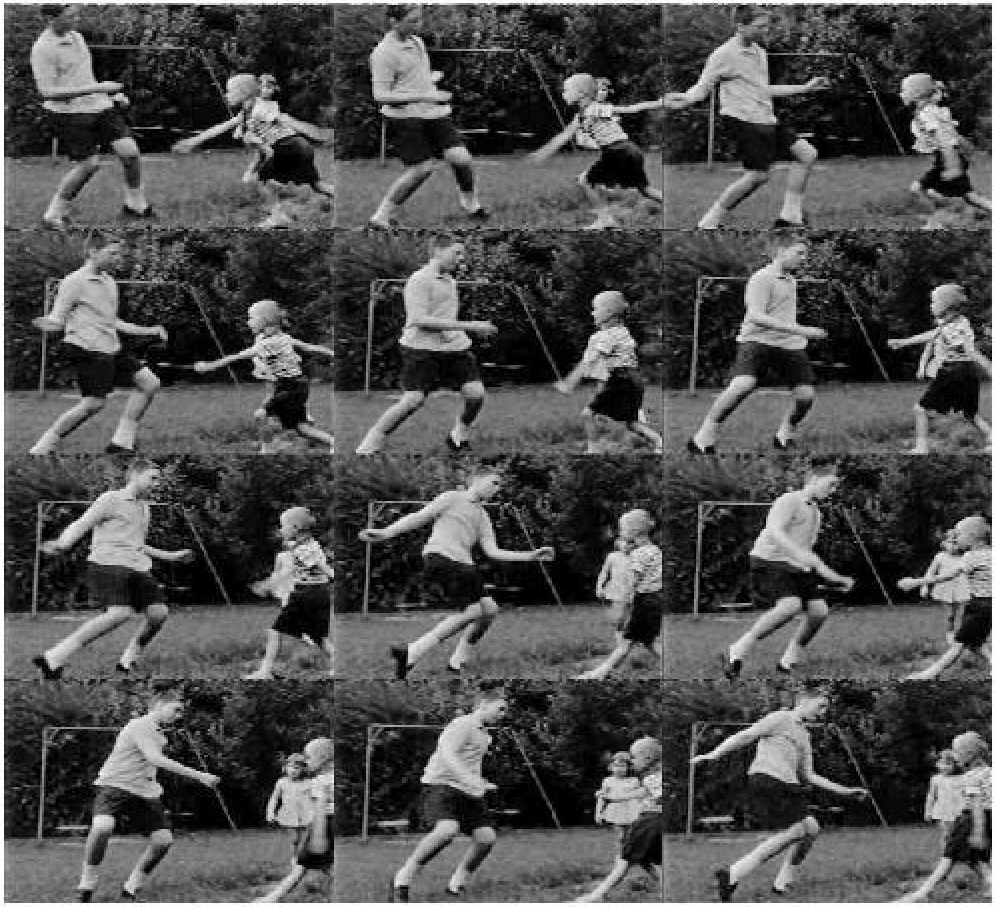
$$N = (npixel \times num\_orientation \times num\_scale)^{body\_parts}$$

$$= (10^6 \times 10^2 \times 5)^9 = 6 \times 10^{78}!$$

Existing Approach

- Controlled setting: background subtraction
- Hand initialization, user assisted tracking
- Grouping based body detection:
  - Forsyth, Finding Naked Body, ECCV’96
  - Felzenszwalb, Huttenlocher, ICCV’99

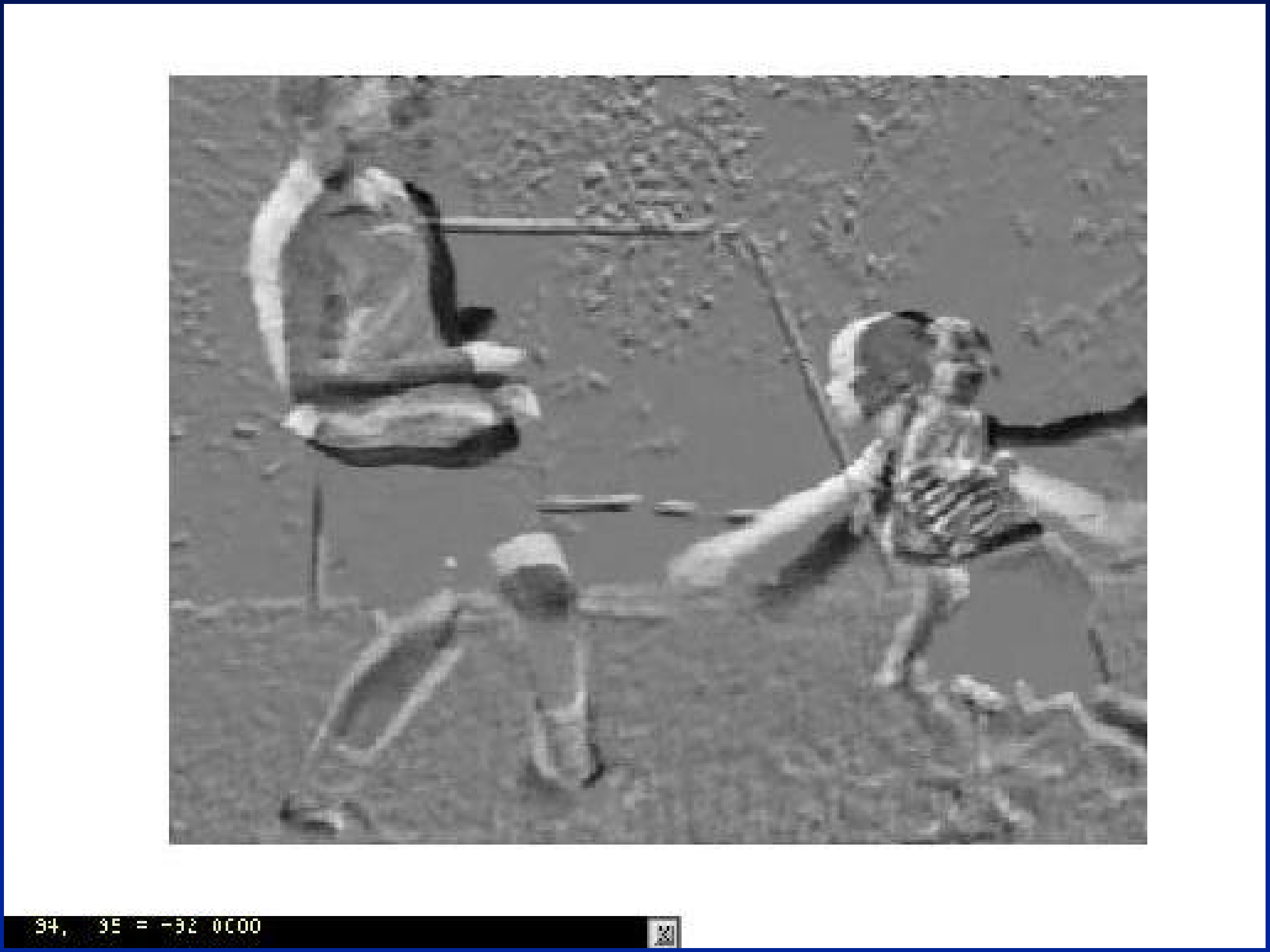
Body finding in a video:



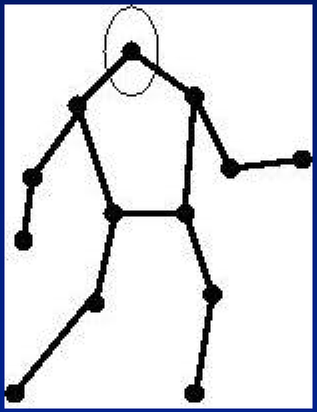
Proposed approach

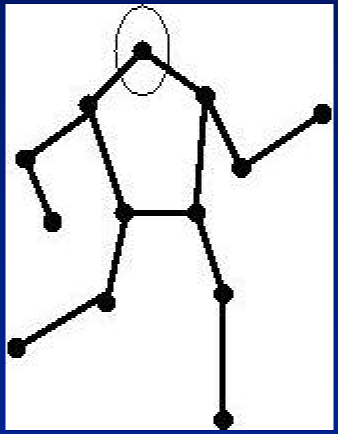
- Using motion information to find possible locations of joints,
- Using MRF inferencing technique for finding the globally optimal body configuration.

Joint location from motion field

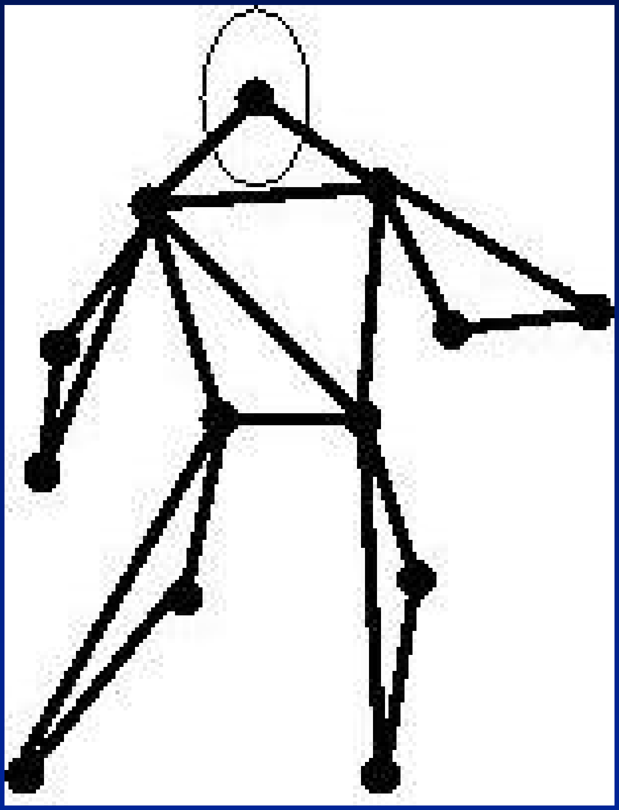


```
94,  95 = -32 0000
```





Body joint tree:

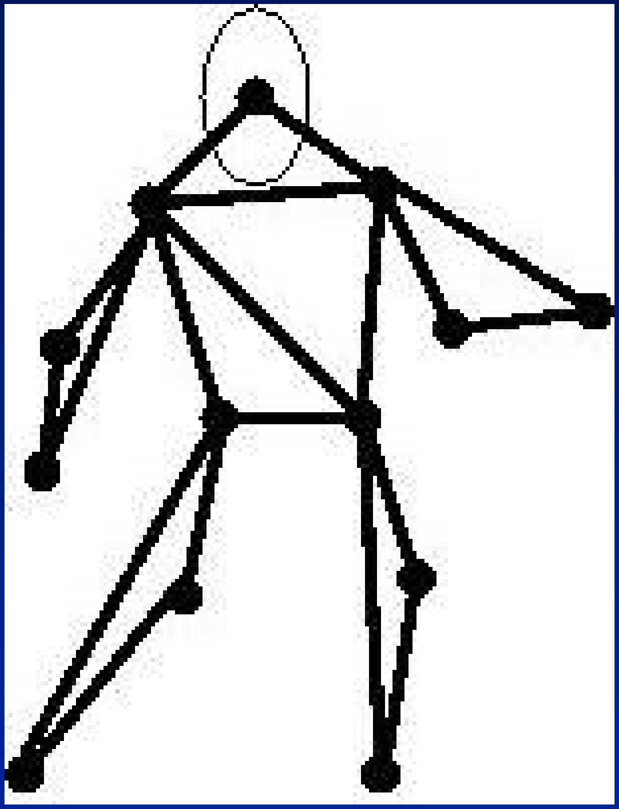


Let $L = (l_{1,} l_2, \ldots, l_n)$, be the assignment of each joints

$$E(L) = \sum_{(l_i, l_j, l_k)} E(l_{i,} l_{j,} l_k) + \sum_{l_i} E(I, l_i),$$

$$P(L) = \frac{1}{Z} e^{-E(L)}$$

Body joint tree:



$$E(L) = \sum_{(l_i, l_j, l_k)} E(l_i, l_j, l_k) + \sum_{l_i} E(I, l_i),$$

$E(I, l_i) =$ how likely l_i is a joint,

$E(l_i, l_j, l_k) =$ how likely is this configurat ion of three joints.

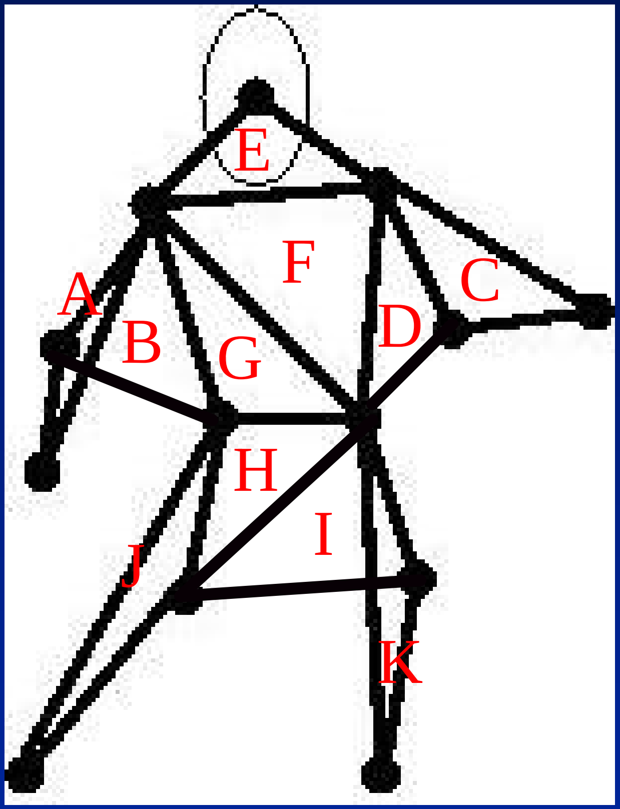
Depends on :

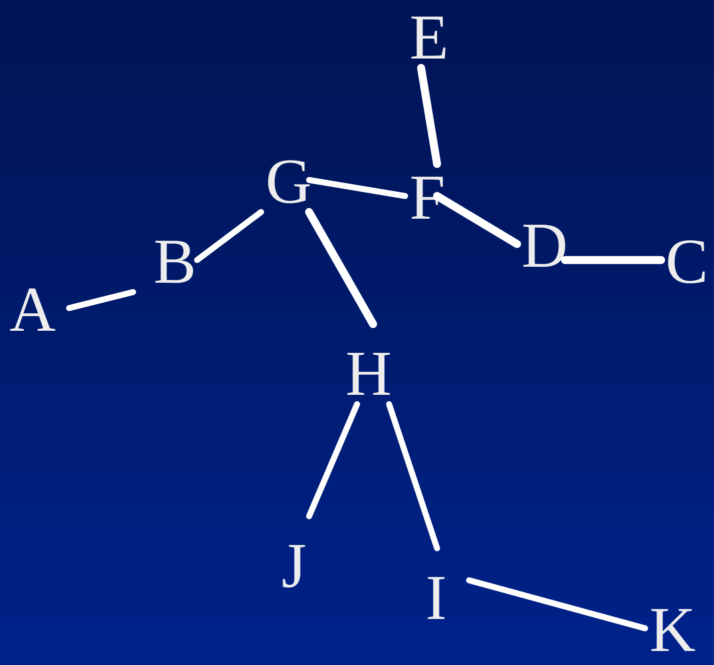
1) geometrica l relationsh ip,

2) relative motion,

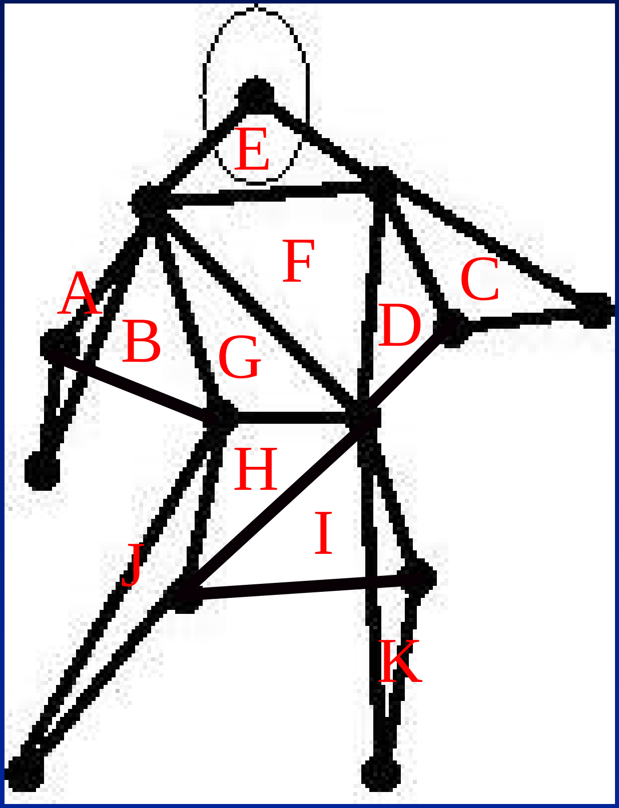
3) image informatio n.

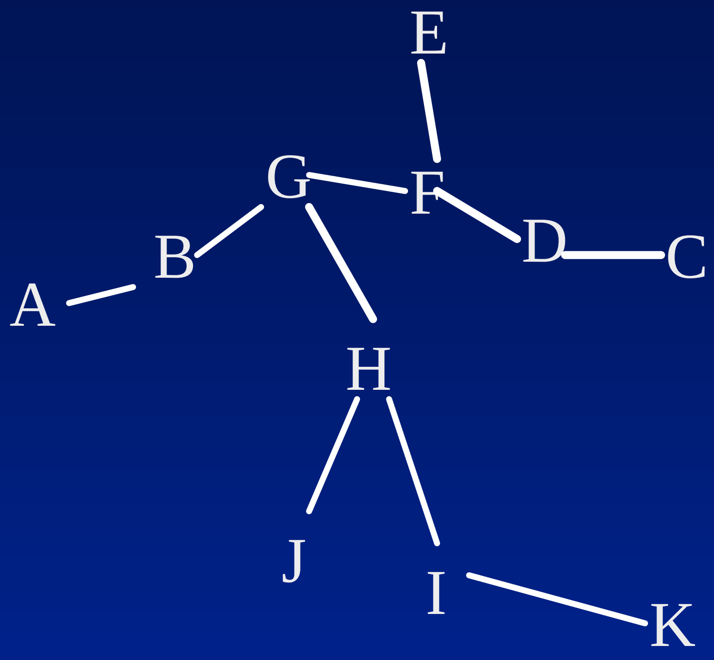
Clique tree





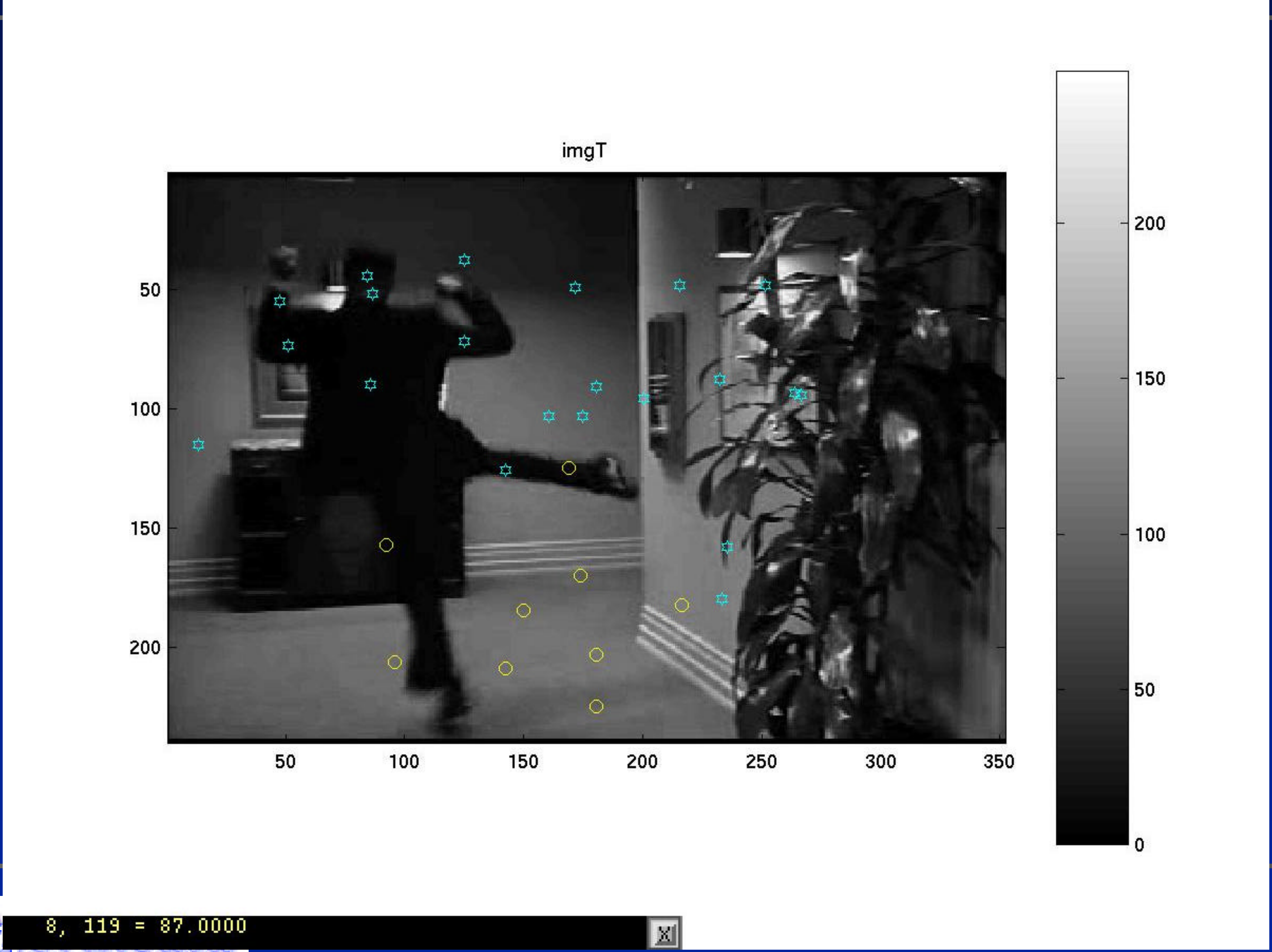
Clique tree





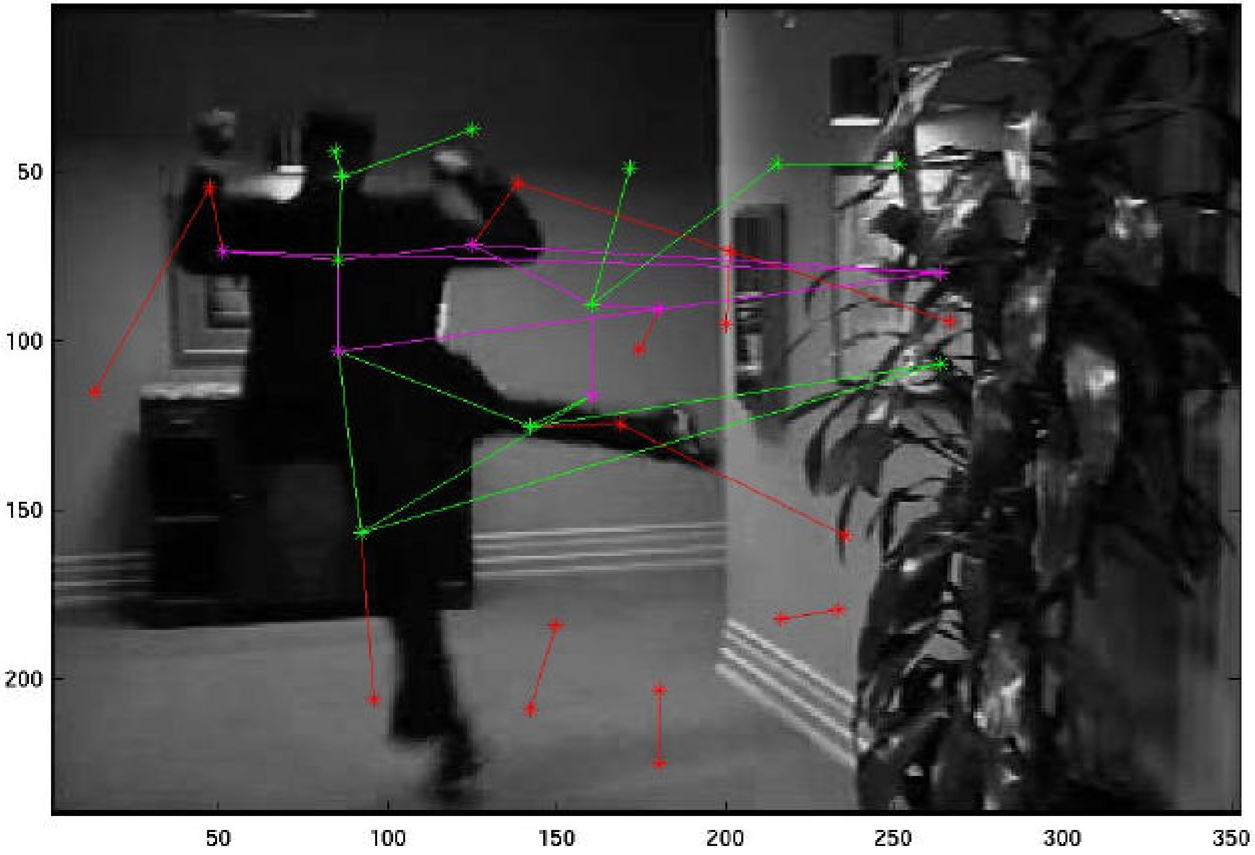
Run Dynamic programming on the clique tree above

Possible body-part locations:

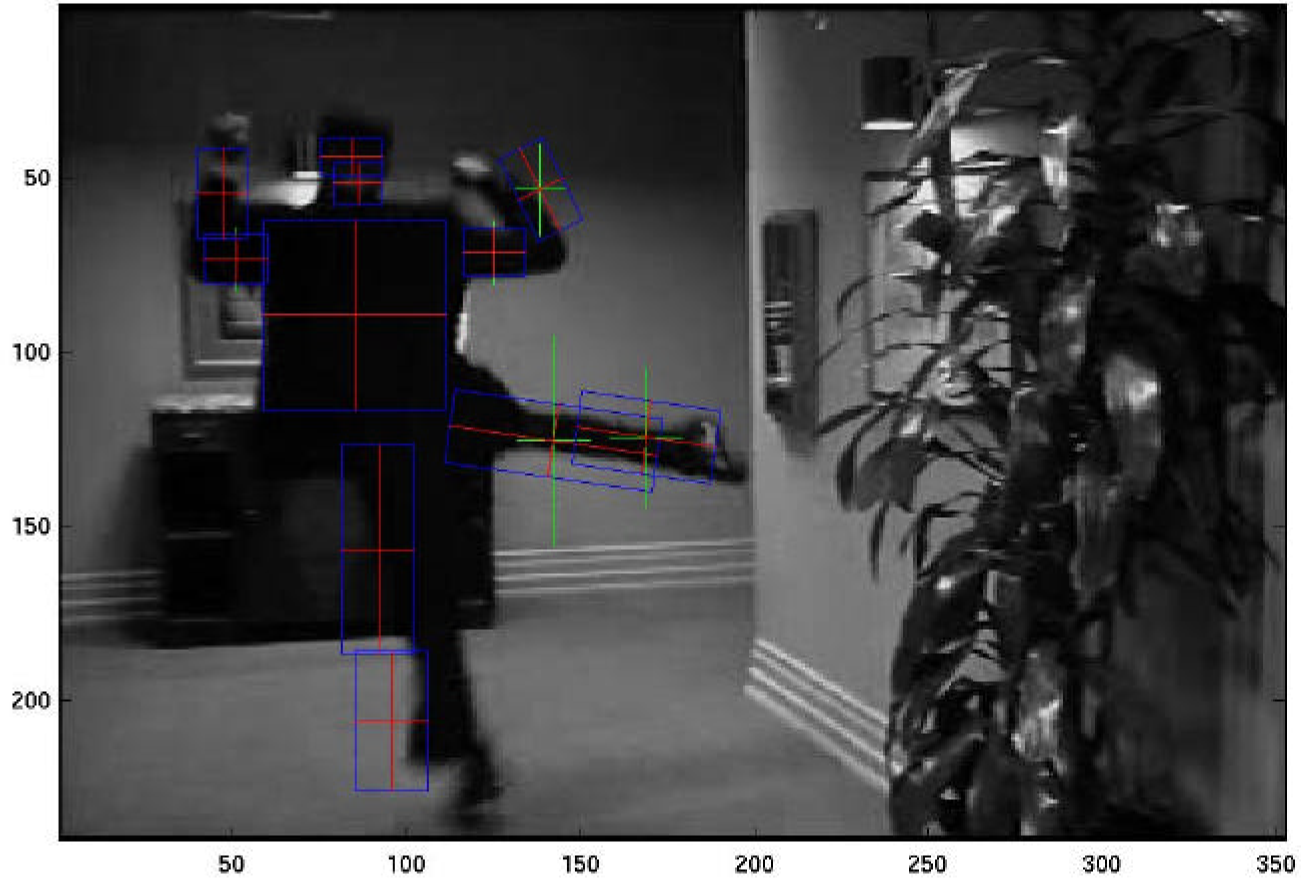


```
8, 119 = 87.0000
```

Some of the possible body configurations:



Optimal configuration:



Scaling Video Image Processing Engine

Specialize for indexing Human and Human Activities

Intelligent Compact Video Representation for Fast trainable query

Data-mining from mapping Between Video and Text

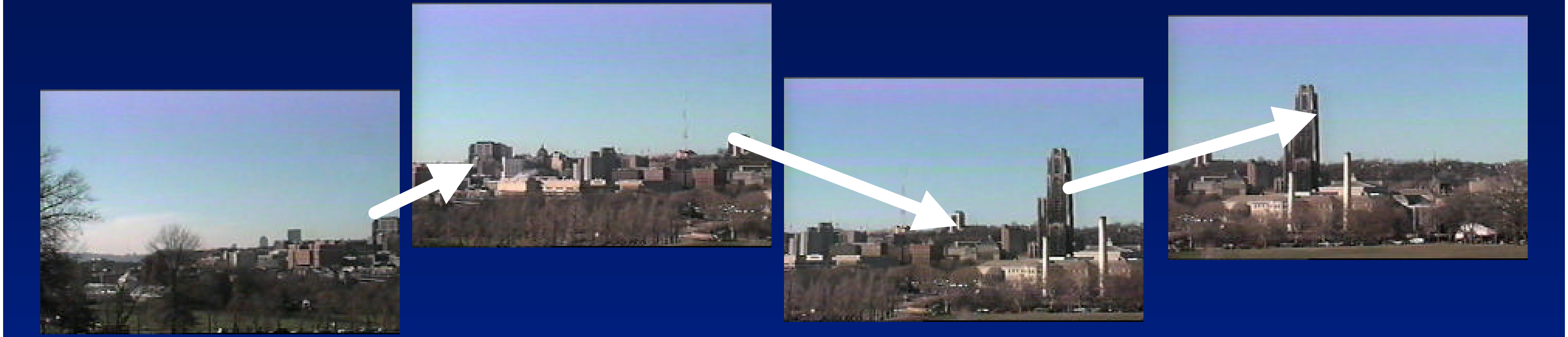
Efficient Representation of Video

Original Video Segment

(3.4Mb)

informedia

Image Registration:



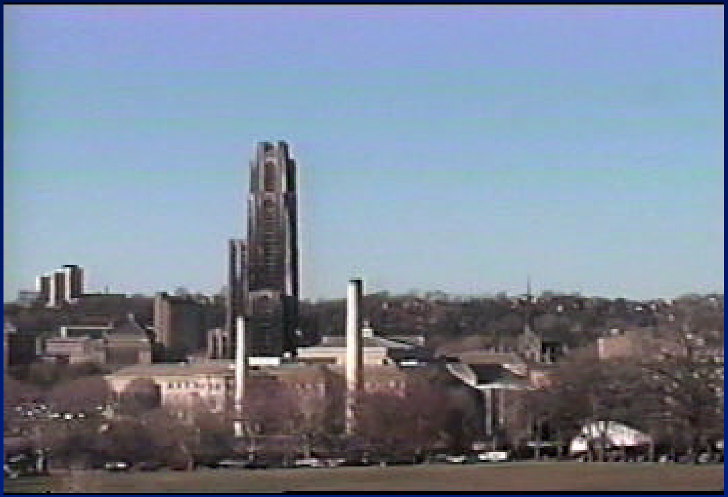
Compute the transformation between image I and J, using affine approximation:

$$J(Ax+D) = I(x)$$

Efficient Representation of Video

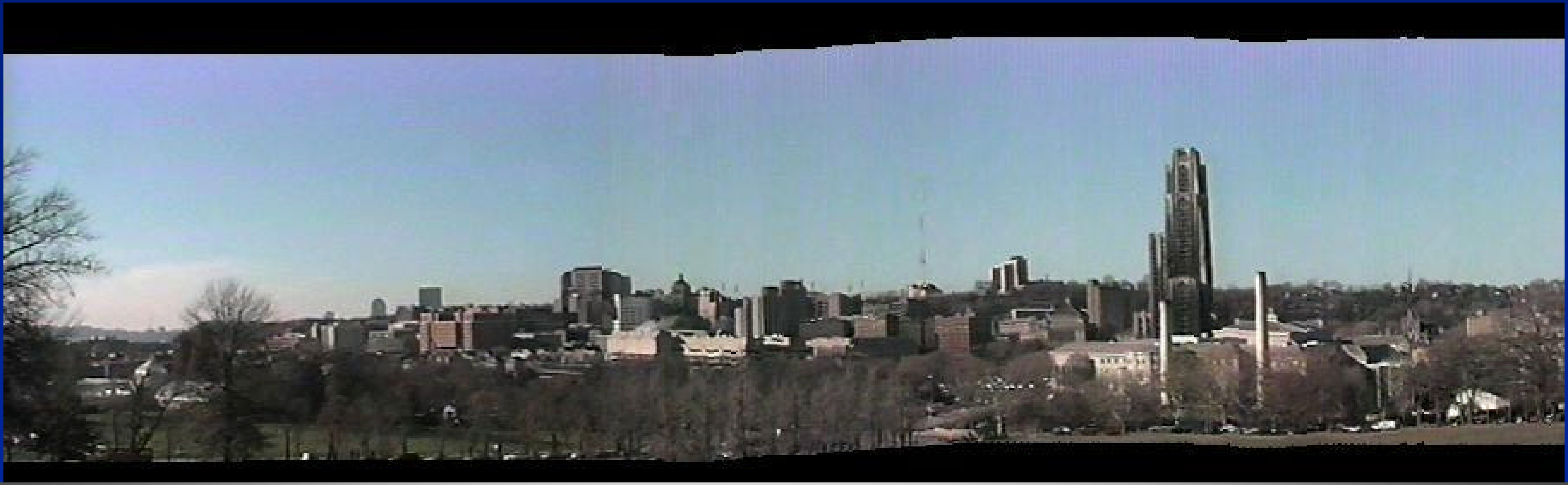
Original Video Segment

(3.4Mb)



Panorama Layer

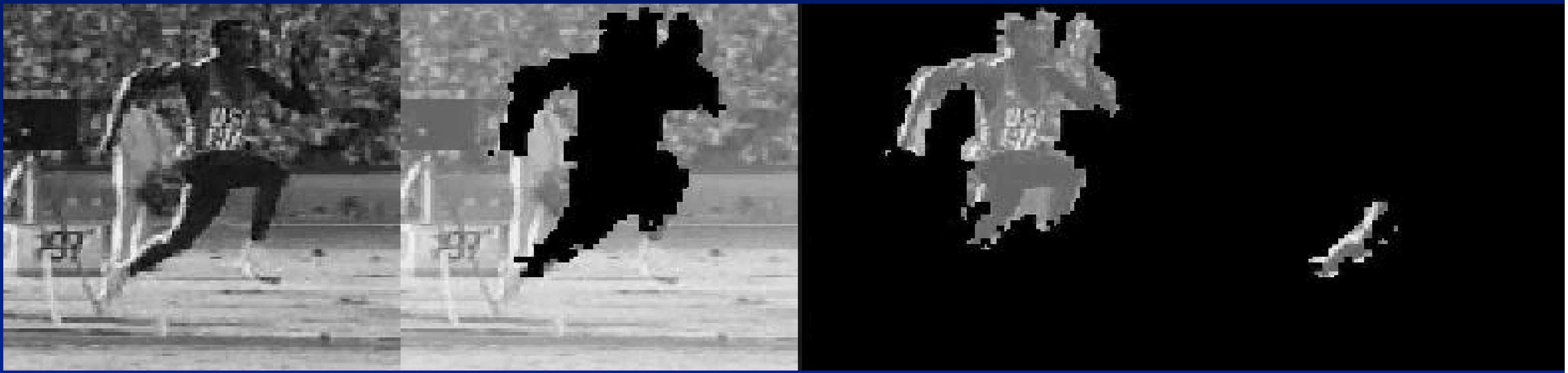
(35Kb)



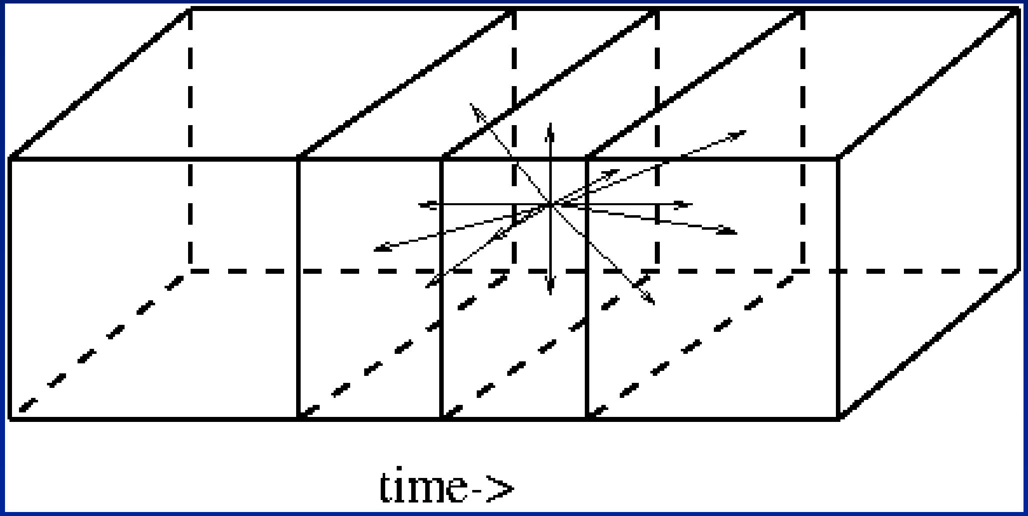
Multiple layers

- Multiple object motion

Motion decoupling



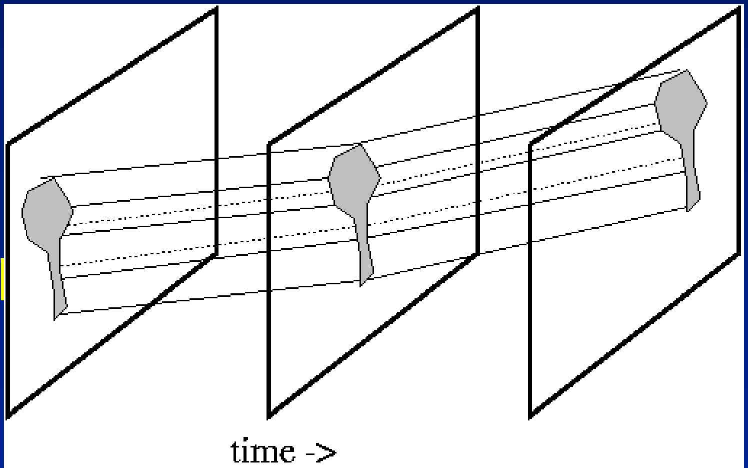
- Networks of spatial-temporal connections: Motion Segmentation with Normalized Cuts



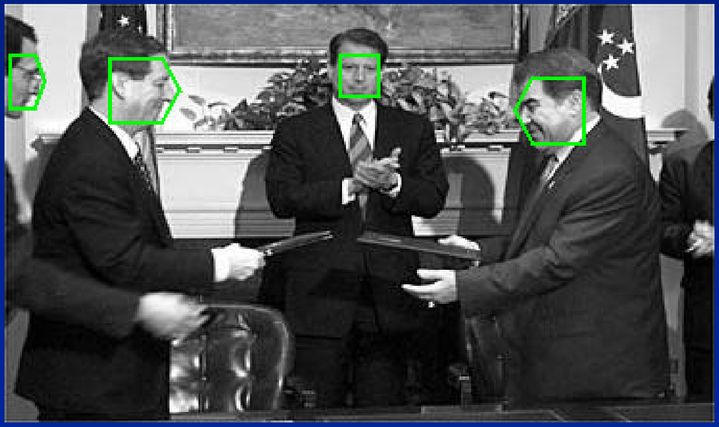
Motion Segmentation with Normalized Cuts

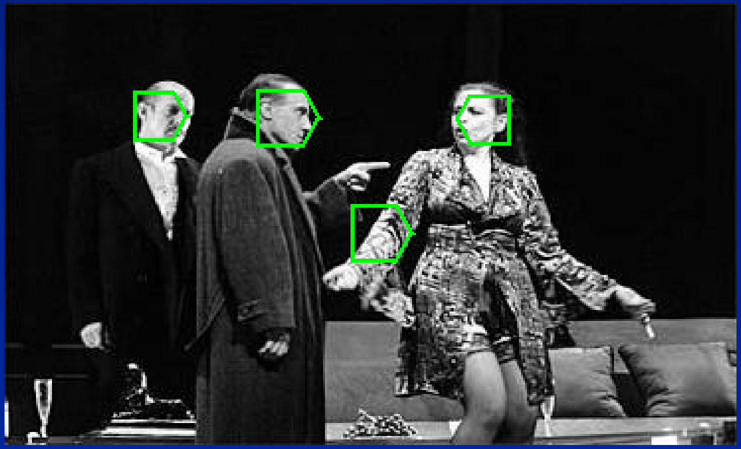
- Motion “proto-volume” in space-time

- Group co



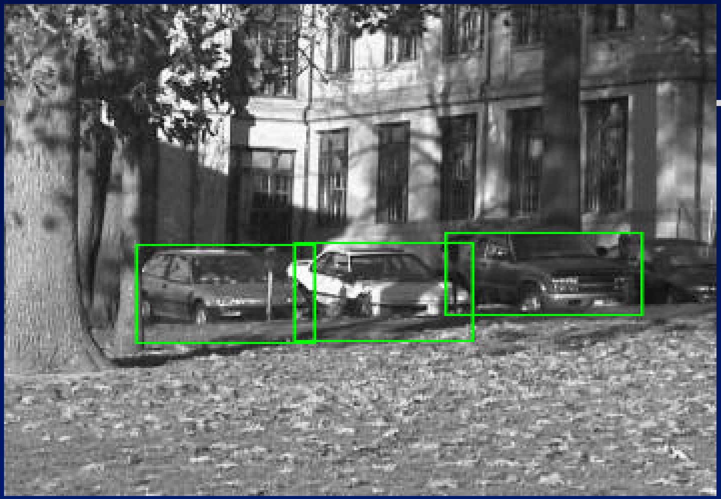
Rapid training of new objects in Video

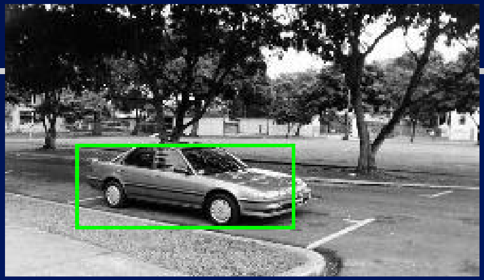


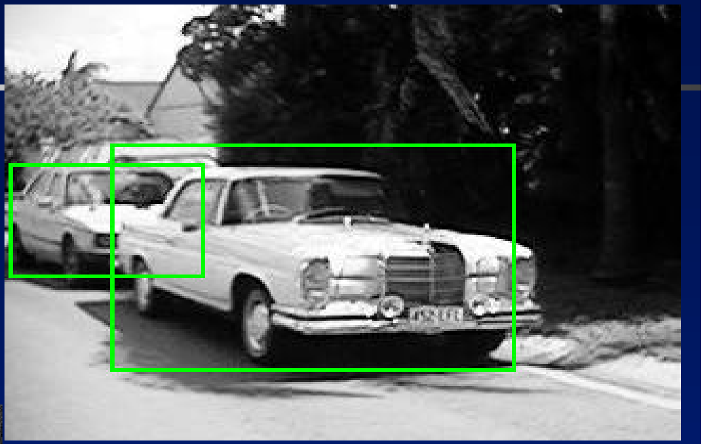


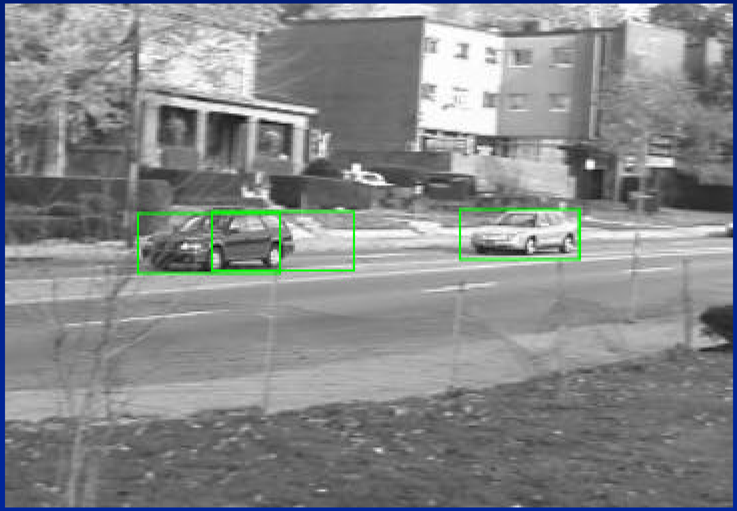
Multi-level object representation/training

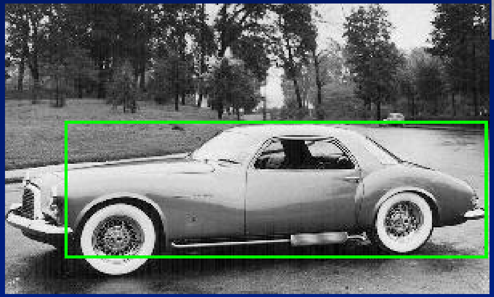
- Representation objects/scene at multiple level of abstraction
  - Low level motion, texture description
  - Mid level object segmentation
  - High level object type and instances

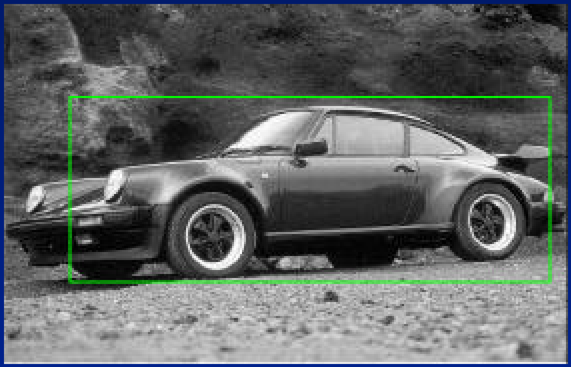


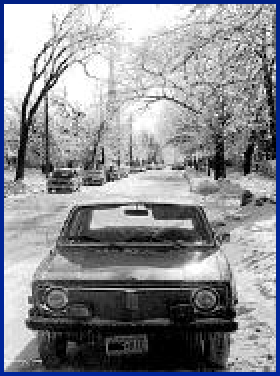


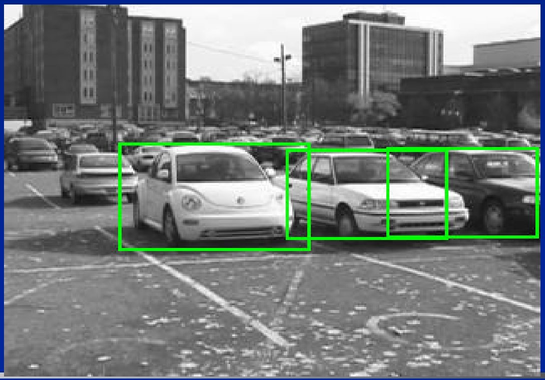


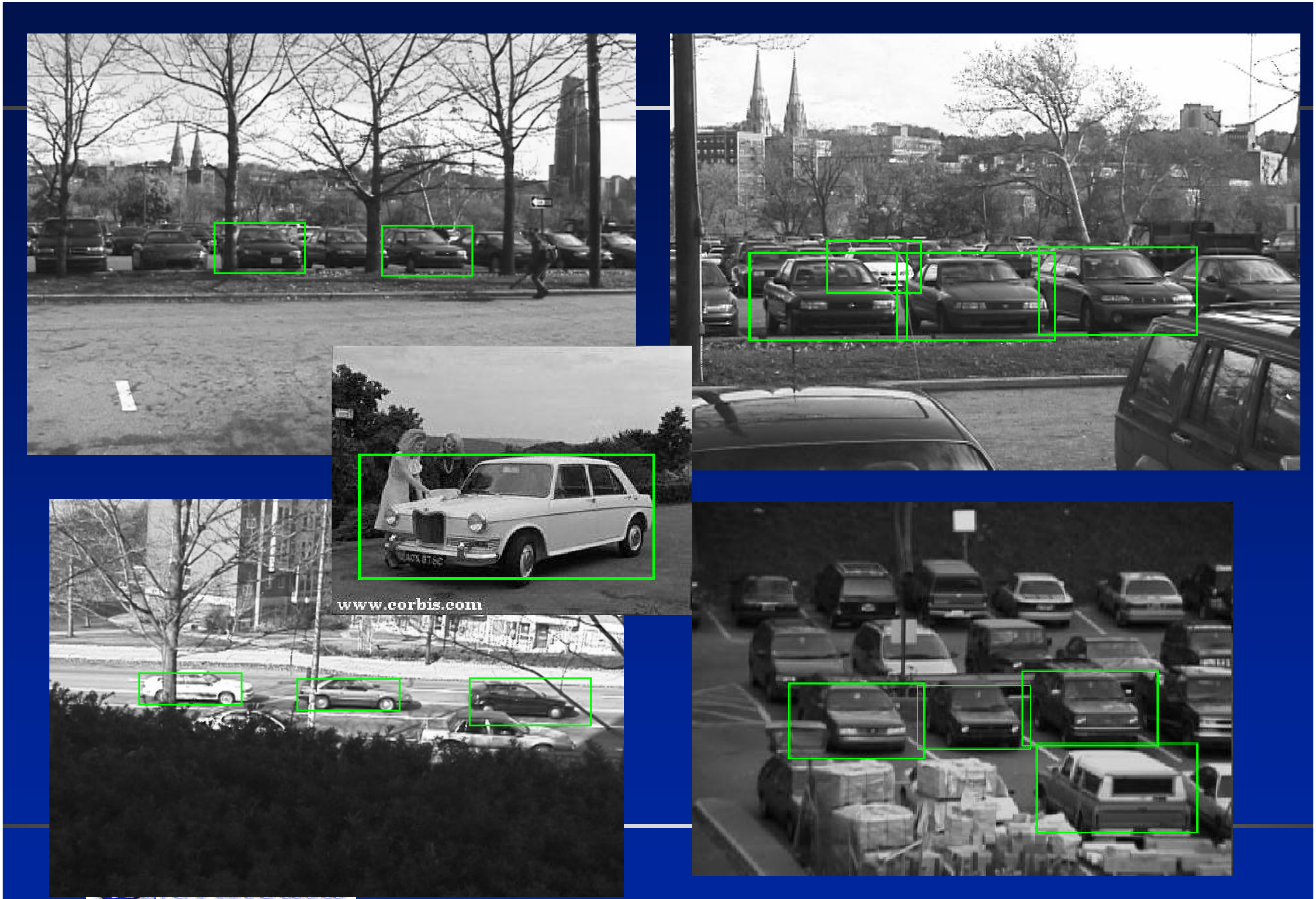




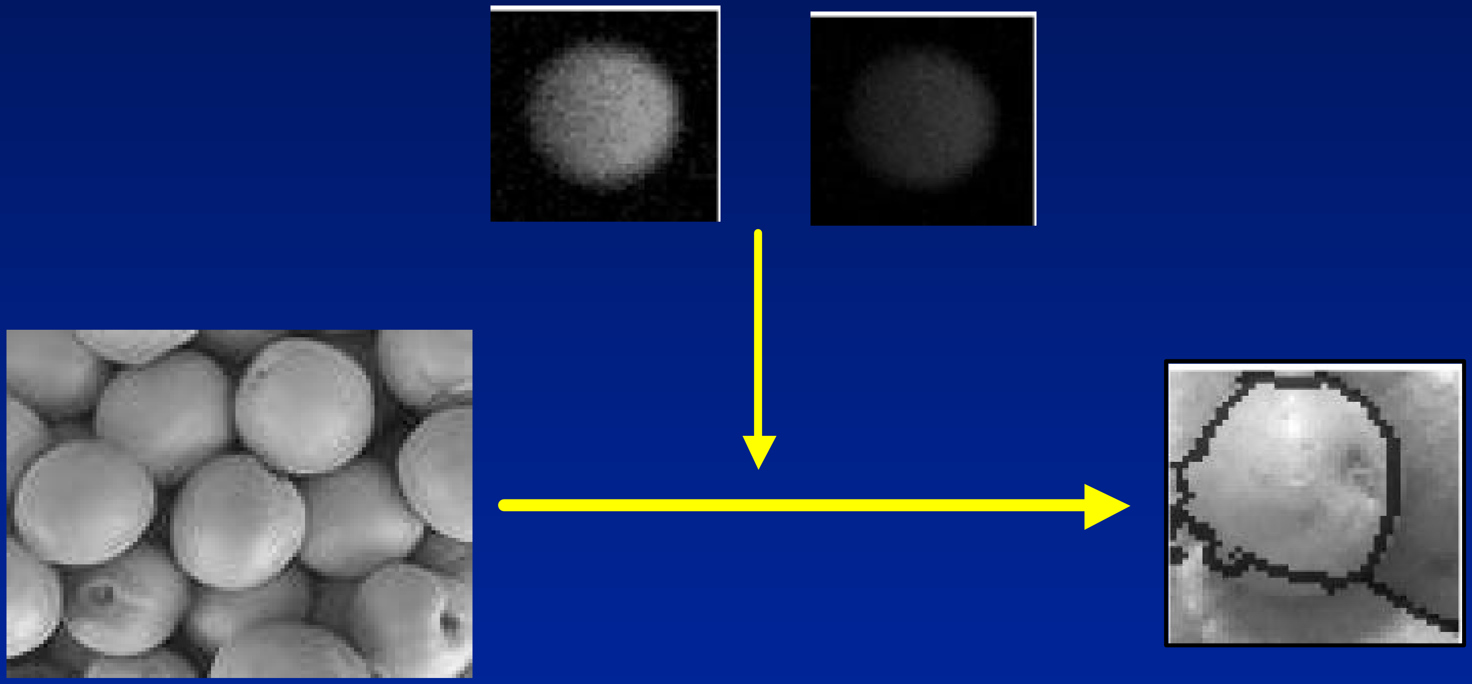




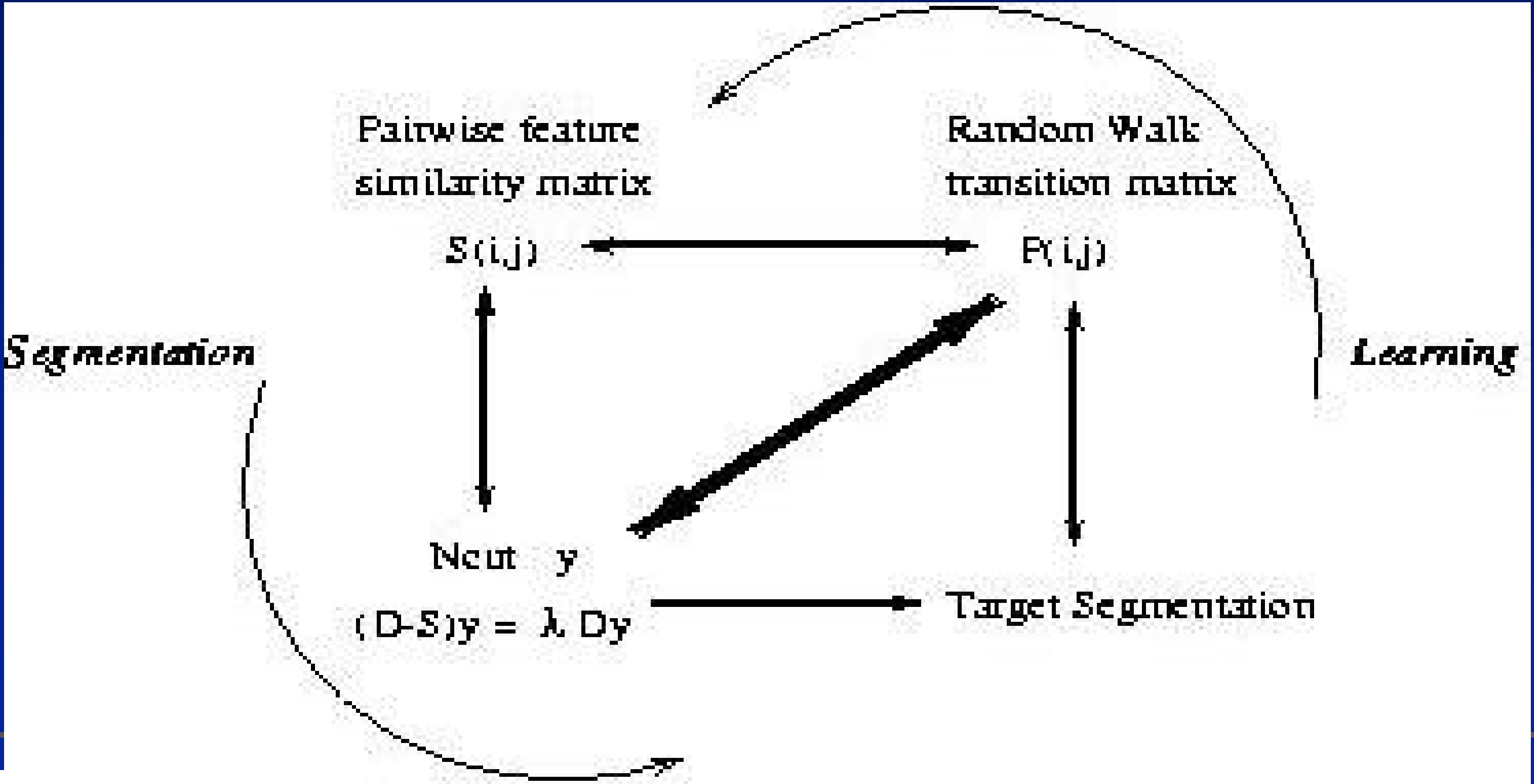



www.corbis.com

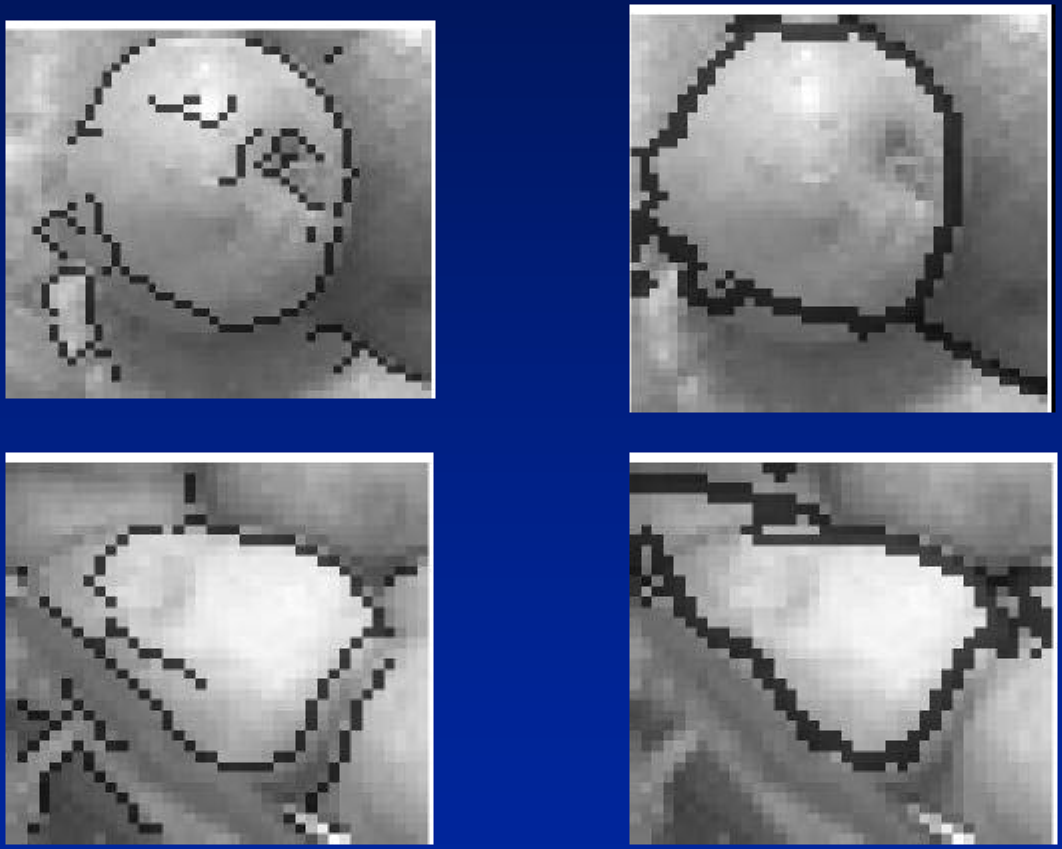
Learning Segmentation



Learning Seg. With Random Walk



Learning Convexity



Scaling Video Image Processing Engine

Specialize for
indexing
Human and Human
Activities

Intelligent Compact
Video Representation for
Fast trainable query

Data-mining from mapping
Between Video and Text

Learning Verbal image query

- Looking for cell phone images

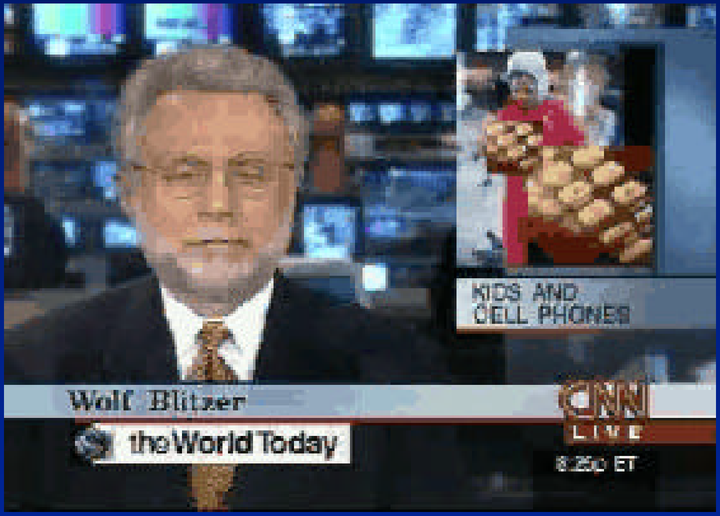


Image learning from text

- Keyword annotation is not sufficient
  - Too many objects
  - Disagreement of keywords

Conclusion

- Move towards more powerful image-query system in Video

Specialize for indexing
Human and Human Activities

Intelligent Compact Video Representation for
Fast trainable query

Data-mining from mapping
Between Video and Text

Conclusion

- Move towards more powerful image-query system in Video

Specialize for indexing
Human and Human Activities

Intelligent Compact Video Representation for
Fast trainable query

Data-mining from Co-relation
Between Video and Text